Possibilistic classifiers for numerical data

Myriam Bounhas · Khaled Mellouli · Henri Prade · Mathieu Serrurier

Abstract Naive Bayesian Classifiers, which rely on independence hypotheses, together with a normality assumption to estimate densities for numerical data, are known for their simplicity and their effectiveness. However, estimating densities, even under the normality assumption, may be problematic in case of poor data. In such a situation, possibility distributions may provide a more faithful representation of these data. Naive Possibilistic Classifiers (NPC), based on possibility theory, have been recently proposed as a counterpart of Bayesian classifiers to deal with classification tasks. There are only few works that treat possibilistic classification and most of existing NPC deal only with categorical attributes. This work focuses on the estimation of possibility distributions for continuous data. In this paper we investigate two kinds of possibilistic classifiers. The first one is derived from classical or flexible Bayesian classifiers by applying a probability–possibility transformation to Gaussian distributions, which introduces some further tolerance in the description of classes. The second one is based on a direct interpretation of data in possibilistic formats that exploit an idea of proximity between data values in different ways, which provides a less constrained representation of them. We show that possibilistic classifiers have a better capability to detect new instances for which the classification is ambiguous than Bayesian classifiers, where probabilities may be poorly estimated and illusorily precise. Moreover, we propose, in this case, an hybrid possibilistic classification approach based on a nearest-neighbour heuristics to improve the accuracy of the proposed possibilistic classifiers when the available information is insufficient to choose between classes. Possibilistic classifiers are compared with classical or flexible Bayesian classifiers on a collection of benchmarks databases. The experiments reported show the interest of possibilistic classifiers. In particular, flexible possibilistic classifiers perform well for data agreeing with the normality assumption, while proximity-based possibilistic classifiers outperform others in the other cases. The hybrid possibilistic classification exhibits a good ability for improving accuracy.

Keywords Naive Possibilistic Classifier · Possibility theory · Proximity · Gaussian distribution · Naive Bayesian Classifier · Numerical data

Communicated by E. Huellermeier.

M. Bounhas (✉) · K. Mellouli
Laboratoire LARODEC, ISG de Tunis, 41 rue de la liberté,
2000 Le Bardo, Tunisia
e-mail: Myriam_Bounhas@yahoo.fr

K. Mellouli
e-mail: Khaled.Mellouli@topnet.tn

H. Prade · M. Serrurier
Institut de Recherche en Informatique de Toulouse (IRIT),
UPS-CNRS, 118 route de Narbonne,
31062 Toulouse Cedex, France
e-mail: Prade@irit.fr

M. Serrurier
e-mail: Serrurier@irit.fr

1 Introduction

Classification is a machine learning technique used to predict class membership for data instances. It consists in searching for algorithms that produce general classifiers from a set of training instances, which constitutes the training phase. The resulting classifier is then used to assign class labels to the testing instances described by a

set of predictor features. This process is usually called testing phase. Classification tasks can be handled by mainly three classes of approaches: those based on empirical risk minimization (decision trees, Quinlan 1986; artificial neural networks, Bishop 1996), approaches based on maximum likelihood estimation (such as Bayesian networks, Pearl 1988), k-nearest neighbours, Cover and Hart 1967) and the ones based on Kolmogorov complexity (Solomonoff 1964). See for instance, Kotsiantis (2007) for a comparative study between these methods.

In this paper we are mainly interested in the second class of methods. Given a new piece of data to classify, this family of approaches seeks to estimate the plausibility of each class with respect to its description (built from the training set of examples), and assigns the class having the highest plausibility value. There are principally two methods: the k-nearest neighbors classifiers and the Naive Bayesian classifiers (NBC). The former, known as lazy learning methods, are based on the principle that an instance to be classified is usually in the proximity of other instances having similar characteristics and that are already labelled. The latter (NBC type) assumes independence of variables (attributes) in the context of classes to estimate the probability distribution on the classes for a given observed data. NBCs are also known for their simplicity, efficiency and small needs in terms of storage space. Moreover NBC perform well, even when making the strong independence assumption which is almost always violated in real datasets (Domingos and Pazzani 2002).

The objective of this paper was to discuss the benefits (and also the limits) of Bayesian-like possibilistic classifiers and to test their feasibility. This work focuses on the classification of data with *numerical* attributes. Three alternatives are commonly considered for handling numerical attributes in an NBC: (i) using a discretization process for continuous attributes and then applying a multinomial probability distribution. It may lead to a loss of information (Yang and Webb 2003) mainly when attributes are discretized in many intervals. However,this method may be effective when the elicitation of the density function turns to be difficult; (ii) assuming normality of the distributions for attributes and estimating the density function using Gaussian densities, or (iii) directly estimating densities in a non-parametric way using kernel density functions.

The study of possibilistic classifiers is motivated by the good performance of NBCs and by the ability of possibility theory (Dubois and Prade 1998) to handle poor data. In spite of the fact that possibility distributions are useful for representing imperfect knowledge, there have been only few works that use Naive Possibilistic Classifiers (Benferhat and Tabia 2008). For this reason, we introduce the Naive Possibilistic Classifiers (NPCs) that are based on the possibilistic counterpart of the Bayesian formula (Dubois and Prade 2000) and the estimation of the possibility distributions.

This work is a fully revised and substantially extended version of a conference paper (Bounhas et al. 2010). In particular, the results on the better handling of ambiguity by possibilistic classifier are new. Moreover in this paper, we also investigate the idea of integrating the possibilistic classifiers with a nearest-neighbors-based heuristic (NNH) in a *hybrid* manner to improve their performances. Indeed, the hybrid classification allows the use of the NNH as an alternative when the main classifier fails to distinguish between classes, i.e. when several classes have very close plausibility estimates. The experiment parts also rely on a larger number of benchmarks (w.r.t. Bounhas et al. 2010) and more evaluation methods.

The paper is structured as follows: in the next section, we give our motivation to the possibilistic classification task and we restate the basic setting of this classification method. In Sect. 3, we study the two kinds of elicitation methods for building possibility distributions: (i) the first one is based on a transformation method from probability to possibility, whereas (ii) the second one makes a direct, fuzzy histogram-based, or possibilistic interpretation of data, taking advantage of the idea of proximity. Section 4 introduces the principle of the hybrid classification. Related works are discussed in Sect. 5. The experimentation results of the proposed approaches are in Sect. 6. The experiments reported show the interest of possibilistic classifiers. In particular, flexible possibilistic classifiers perform well for data agreeing with the normality assumption, while proximity-based possibilistic classifiers outperform other classifiers in the other cases. Moreover, the hybrid classification contributes to significantly improve the accuracy of possibilistic classifier. Finally, Sect. 7 concludes and suggests some directions for future work.

2 General setting of possibilistic classification

We first recall some basics of possibility theory and then present the possibilistic classification viewed as a possibilistic version of the Bayes rule. In the following we also motivate the potential interest of possibility theory in classification.

2.1 Basic notions of possibility theory

Possibility theory (Dubois and Prade 1988) has been introduced by Zadeh (1978). It handles epistemic uncertainty in a qualitative or quantitative way. In particular, possibility theory is suitable for the representation of imprecise information. For a more complete introduction to possibility theory, see (Dubois and Prade 1998).

Possibility theory is based on *possibility distributions*. Given a universe of discourse $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$, a possibility distribution π is a function that associates with each element ω_i from the universe of discourse Ω a value in a bounded and linearly ordered valuation set $(L, <)$. This value is called a *possibility degree*. This scale may be quantitative, or qualitative when only the ordering between the degrees makes sense. In this paper, possibility degrees have a *quantitative* reading and L is taken as the unit interval [0, 1]. A possibility distribution is used as an elastic constraint that restricts the more or less possible values of a single-valued variable.

Possibility distributions have two types of quantitative interpretations. The first one, that is related to fuzzy set theory, is the description of gradual properties. For instance, the definition of linguistic expressions such that "long", "old" or "expensive" does not refer to a specific value, but to a set of possible values in a specific context. For instance, a possibility distribution may describe the concept "expensive" for an house in a particular area. In such a case, each price will be associated with a possibility degree which quantifies how much this price is typical with respect to the concept "expensive". Assigned to events, possibility degrees can also represent *plausibility* which reflects the belief degree of the expert that a certain event will occur. In this scope, a possibility distribution is viewed as a family of probability distributions (see Dubois 2006 for an overview). Thus, a possibility distribution π represents the family of the probability distributions for which the measure of each subset of Ω is bounded by its necessity and its possibility measures.

By convention, $\pi(\omega_i) = 1$ means that it is fully possible that ω_i is the value of the variable. Note that distinct value ω_i, ω_j may be such that $\pi(\omega_i) = 1 = \pi(\omega_j)$. $\pi(\omega_i = 0$ means that ω_i is impossible as the value of the variable. Thanks to the use of the interval [0, 1], intermediary degrees of possibility can be assessed, which enable us to acknowledge that some values are more possible than others.

In possibility theory, different important particular cases of knowledge situation can be represented:

- Complete knowledge: $\forall\omega_i,\ \pi(\omega_i) = 1$ and $\forall\omega_i \neq \omega_j$, $\pi(\omega_j) = 0$.
- Partial ignorance: $\forall\omega_i \in A \subseteq \Omega, \pi(\omega_i) = 1, \forall\omega_i \notin A$, $\pi(\omega_i) = 0$.
- Total ignorance: $\forall\omega_i \in \Omega, \pi(\omega_i) = 1$ (all values in Ω are possible).

A possibility distribution π on Ωenables events to be qualified in terms of their plausibility and their certainty, by means of two dual possibility and necessity measures that are, respectively, defined for an event $A \subseteq 2^\Omega$ by the following formulas:

$$\Pi(A) = \max_{\omega \in A} \pi(\omega)$$

$$N(A) = \min_{\omega \notin A}(1 - \pi(\omega)) = 1 - \Pi(\overline{A})$$

$\Pi(A)$ evaluates to what extent A is consistent with our knowledge represented by π. Indeed, the evaluation provided by $\Pi(A)$ corresponds to a degree of *non-emptiness of the intersection* of the classical subset A with the fuzzy set having π as membership function. Moreover, $N(A)$ evaluates to what extent A is certainly implied by our knowledge since it is a degree of *inclusion* of the fuzzy set corresponding to π into the subset A.

Quantitative possibility distributions can represent, or more generally approximate, a family of probability measures (Dubois and Prade 1992). Indeed, a possibility measure Π can be viewed as an upper bound of a probability measure and associated with the family of probability measures defined by

$$\mathcal{P}(\Pi) = \{P \text{ s.t. } \forall A, \Pi(A) \geq P(A)\}.$$

Thanks to the duality between Π and N and the auto-duality of P ($P(A) = 1 - P(\overline{A})$), it is clear that

$$\forall P \in \mathcal{P}(\Pi),\ \forall A, \Pi(A) \geq P(A) \geq N(A).$$

This is the starting point for defining a probability–possibility transform. The width of the gap between $N(A)$ and $\Pi(A)$ evaluates the amount of ignorance about A since it corresponds to the interval containing the imprecisely known probability. Thus, possibility distributions can in particular represent precise or imprecise information (representable by classical subsets) as well as complete ignorance. The possibilistic representation of complete ignorance should not be confused with a uniform probability distribution. Indeed, with the above representation, we have $\Pi(A) = 1$ for any non-empty event A, and $N(A) = 0$ for any event A different from Ω, while a uniform probability distribution on a universe with more than two elements associates non-trivial events with a probability degree strictly between 0 and 1, which sounds paradoxical for a situation of complete ignorance. Possibility theory is particularly suited for representing situations of partial or complete ignorance (see Dubois 2006; Dubois and Prade 2009, for detailed comparative discussions between probability and possibility).

2.2 Conditional possibility and possibilistic Bayesian rule

Conditioning in possibility theory is defined through a counterpart of Bayes rule, namely

$$\Pi(A \cap B) = \Pi(A|B) * \Pi(B)$$

It has been shown that there are only two basic choices for $*$, either minimum or the product (Dubois and Prade 1990).

The min operator is suitable in the qualitative possibility theory setting, while the product should be used in quantitative possibility theory (De Cooman 1997). Quantitative possibilistic conditioning can be viewed as a particular case of Dempster's rule of conditioning since possibility measures are special cases of plausibility functions (Shafer 1976).

Thus, possibilistic conditioning corresponds to revising an initial possibility distribution π, when a new information $B \subseteq \Omega$ is now available. In the quantitative setting we have

$$\pi(a \mid B) = \begin{cases} \frac{\pi(a)}{\Pi(B)} & \text{if } a \in B \\ 0 & \text{otherwise} \end{cases}.$$

2.3 Naive Bayesian possibilistic classification

The idea of applying possibility theory to classification parallels the use of probabilities in Bayesian classifiers (see the "Appendix" for a reminder). Probability distributions used in NBCs are usually built by assuming that numerical attributes are normally distributed around their mean. Even if a normal distribution is appropriate, identifying it exactly from a sample of data is especially questionable when data are poor. When normality assumption is violated, Gaussian kernels can be used for approximating any type of distributions. Then, it is required to assess many parameters, a task that may be not compatible with poor data. The problem of the precise estimation of probability distributions for NBCs is important for the exact computation of the probability distribution over the classes. However, due to the use of the product for combining probability values (which are often small), the errors on probability estimations may have a significant effect on the final estimation. This contrasts with possibility distributions which are less sensitive to imprecise estimation for several reasons. Indeed, a possibility distribution may be viewed as representing a family of probability distributions corresponding to imprecise probabilities, which sound more reasonable in case of poor data. Moreover, we no longer need to assume a particular shape of probability distribution in this possibilistic approximation process.

As in the case of Bayesian classification, possibilistic classification is based on the possibilistic version of the Bayes theorem. Given a new vector $\{a_1, \ldots, a_M\}$ of n observed variables $A_1, \ldots, A_M$ and the set of classes $C = \{c_1, \ldots, c_C\}$, the classification problem consists in estimating a possibility distribution on classes and in choosing the class with the highest possibility for the vector X in this quantitative setting, i.e.

$$\Pi(c_j|a_1, \ldots, a_M) = \frac{\pi(a_1, \ldots, a_M|c_j) * \Pi(c_j)}{\pi(a_1, \ldots, a_M)} \quad (1)$$

In formula (1), the quantitative component of possibilistic classification involves *prior* possibility distribution relative to the class and the input vector. Note that the term $\pi(a_1, \ldots, a_M)$ is a normalization factor and it is the same over all class values. In this work, we assume that there is no a priori knowledge about classes and the input vector to classify (thus $\pi(c_j) = 1$ and $\pi(a_1, \ldots, a_M) = 1$)). Moreover, as in naive Bayesian classification, naive possibilistic classification assumes that variables A_i are independent in the context of classes (Ben Amor et al. 2002).

Assuming attribute independence, the plausibility of each class for a given instance is computed as

$$\pi(c_j|a_1, \ldots, a_M) = \prod_{i=1}^{M} \pi(a_i|c_j) = \pi(a_1|c_j) * \cdots * \pi(a_M|c_j) \quad (2)$$

where conditional possibilities $\pi(a_i|c_j)$ in formula (2) represent to what extent a_i is a possible value for the attribute A_i in the presence of the class c_j. As in the case of the conditioning rule, * may be chosen as the min or the product operator (min corresponds to complete logical independence, while the use of the product makes partially possible values jointly less possible). In a product-based setting, a given instance is assigned to the most plausible class c^*:

$$c^* = \arg\max_{c_j} \prod_{i=1}^{M} \Pi(a_i|c_j) \quad (3)$$

It is worth noticing that formula (2) has a set-theoretical reading. Namely, when the possibility distributions take only the values 0 and 1, the formula (2) amounts to express that an instance may be possibly classified in c_j inasmuch as the attribute value of this instance is compatible with this class given the available observations. Thus, possibilistic classification may be viewed as an intermediate between Bayesian probabilistic classification and a purely set-based classifier (such classifiers use as distributions the convex hull for each attribute of the data values to identify classes, usually leading to too many multiple classifications).

3 Elicitation of the possibility distributions

In this section, we describe several methods for building possibility distributions from data belonging to continuous domains. We consider two families of approaches: the first one is based on a probability–possibility transformation method (Dubois et al. 1993, 2004; Yamada 2001). The second one is based on a direct possibilistic interpretation of data taking advantage of the idea of proximity.

In this approach and for all the rest of this work, all attribute values a_i's are normalized as follows:

$$a_{i_n} = \frac{a_i - \min(a_i)}{\max(a_i) - \min(a_i)}, \quad (4)$$

min and max are functions giving respectively the minimum and the maximum value of the attribute a_i over the training set. For the sake of simplicity we use in the rest of the paper only normalized attribute values, e.g., every attribute value a_i refers to the corresponding a_{i_n}.

3.1 Probability to possibility transformation method

The transformation from probability to possibility distributions (Dubois et al. 2004), which relies on the view of possibilities as upper bounds of probabilities, has been extended to continuous universes. It yields the most restrictive possibility distribution that is comonotone with the probability distribution and that provides an upper bound on the probability of any event.

3.1.1 Principle

Let us recall the principle underlying the transformation from probability distribution p to possibility distribution π* proposed in Dubois et al. (1993, 2004). The resulting possibility distribution should satisfy the following properties:

- Possibility–probability consistency: For any probability density p, the possibility distribution π* is consistent with p, that is $\forall A, \Pi^*(A) \geq P(A)$, with Π^* and P being the possibility and probability measures associated with π* and p, respectively.
- Comonotony of distributions: $\pi(x) > \pi(x')$ if and only if $p(x) > p(x')$.

The rationale behind this transformation is that given a probability p, one tries to preserve as much information as possible. This leads to select the most specific element in the set $\mathcal{PI}(P) = \{\Pi\colon \Pi \geq P\}$ of possibility measures dominating P such that $\pi(x) > \pi(x')$ iff $p(x) > p(x')$.

Dubois et al. (1993) have justified a probability–possibility transformation method in the continuous case in terms of confidence intervals (with level ranging from 0 to 1) built around a nominal value which is the mode. It generalizes a previously proposed method for the discrete case (Dubois et al. 1993). In this context, densities are assumed to be symmetric with unique mode. Then, the mode is equal to the mean and the median. A confidence interval I_α represents the smallest range of values that is believed to include the "true" value of the considered variable, with a fixed probability α. Its confidence level is $P(I_\alpha) = \alpha$, $1 - P(I_\alpha)$ is the risk level, that is, the probability for the real value to be outside the interval. It leads to build the following possibility distribution π* in the continuous case:

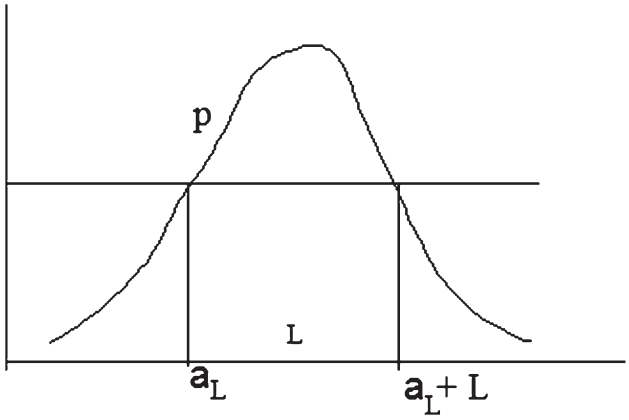


Fig. 1 Confidence Interval I_L for a given a_L

$$\pi^*(x) = sup\{1 - P(I_\alpha), x \in I_\alpha\}, \quad (5)$$

where I_α is the α-confidence interval. It has been shown (Dubois et al. 2004) in case of a unimodal distribution that the associated possibility distribution is such that

$$\forall L > 0,\ \pi(a_L) = \pi(a_L + L) = 1 - P(I_L), \quad (6)$$

where I_L is the smallest confidence interval, of length L, that contains a_L, here assumed to be smaller than the mode (see Fig. 1).

In this section, we propose two elicitation approaches based on the probability to possibility transformation method. We apply this transformation method to a Gaussian distribution, which leads to two classifiers called Naive Possibilistic Classifier and Flexible Naive Possibilistic Classifier, respectively.

3.1.2 Gaussian density-based transformation: the Naive Possibilistic Classifier (NPC)

We have to find a possibility distribution over a training set which is the most specific representation for the numerical data. According to (6), $\pi(a_i|c_j)$ can be estimated by $1 - P(Ia_i|c_j)$ where Ia_i is the confidence interval as previously said. The main question is to estimate such confidence interval for each attribute a_i belonging to the class c_j.

For the NPC, we assume that each attribute value a_i is a random variable which is normally distributed over the class c_j. Thus for each class c_j, a Gaussian distribution $g_{ij} = g(a_i,\ \mu_{ij},\ \sigma_{ij})$ should be given. For such Gaussian, μ_{ij} is the mean of the variable a_i for the class c_j and σ_{ij} is its standard deviation for the same class. They are estimated from the training dataset.

If Ia_i is the confidence interval centered at μ_{ij}, its probability $P(Ia_i|c_j)$ can be estimated by

$$P(I_{a_i}|c_j) = 2 * G(a_i, \mu_{ij}, \sigma_{ij}) - 1, \quad (7)$$

where G is a Gaussian cumulative distribution easily evaluated using the table of the standard normal distribution. Thus, we estimate $\pi(a_i\,|c_j)$ using the following formula:

$$\begin{aligned}\pi(a_i|c_j) &= 1 - (2 * G(a_i, \mu_{ij}, \sigma_{ij}) - 1) \\ &= 2 * (1 - G(a_i, \mu_{ij}, \sigma_{ij})). \end{aligned} \quad (8)$$

Hence, in the training phase we should simply calculate the mean μ_{ij} and the standard deviation σ_{ij} for each attribute a_i of instances belonging the class c_j.

3.1.3 *Kernel density-based transformation: the Flexible Naive Possibilistic Classifier (FNPC)*

The FNPC is mainly based on the FNBC as presented in the "Appendix". For this classifier, the building procedure is reduced to the calculation of the standard deviation σ. The FNPC is the same as the NPC in all respects, except that it uses a different method for density estimation. Instead of using a single Gaussian to estimate each continuous attribute, we investigate kernel density estimation as in the FNBC.

It is proved in John and Langley (1995) that classifiers based on kernel estimation are more accurate than Gaussian-based estimation to fit non-Gaussian densities. The idea of estimation based on Gaussian kernels (see "Appendix") may be adapted in the possibilistic context in the spirit of formula (8). Haouari et al. (2009) have justified the use of the arithmetic mean function to estimate a possibility distribution for an attribute given the class when dealing with n individual possibilities over the training set.

If we just consider that we have to combine possibility measures (forgetting how they have been obtained) the natural way to do it is to use a weighted maximum operator (Dubois and Prade 1990). However, our problem, as announced in Haouari et al. (2009), is closer to being an estimation task than a fusion because training instances come from a single random source and not from n independent sources of information. Besides, the authors show that the arithmetic mean function satisfies the three necessary properties allowing it to be an estimation function: *idempotency*, *commutability* and *monotonicity* (see Haouari et al. 2009, for details).

$$\pi(a_i|c_j) = \frac{1}{N_j}\sum_{k=1}^{N_j} \pi(a_i, c_{jk}). \quad (9)$$

with

$$\pi(a_i, c_{jk}) = 2 * (1 - G(a_i, \mu_{ik}, \sigma)). \quad (10)$$

where k ranges over the N_j instances of the training set in class c_j and $\mu_{ik} = a_{ik}$.

Various rules are used in the statistical literature for setting the kernel width σ. John and Langley (1995) have proved that the Flexible Bayes classifier is strongly consistent if the kernel density estimate satisfies the theorem of *strong consistency* (Devroye 1983). In this theorem, two necessary conditions for the width σ of the kernel density estimate must be satisfied: (i) $\sigma \to 0$ as $n \to \infty$ and (ii) $n\sigma \to \infty$ as $n \to \infty$.

In this paper and for all distributions, the standard deviation is estimated by:

$$\sigma = \frac{1}{\sqrt{N}}. \quad (11)$$

Both σ estimators in formula (11) and (26) (in the "Appendix") satisfy the two conditions of strong consistency theorem. However, we empirically choose to use the estimator in formula (11) due to its better performance in experimentations. It may be due to the fact that the density estimated became increasingly local when we consider all training instances (N) instead of considering only those belonging to a specific class (N_j) when estimating σ. The intuition behind this choice is that this estimator will contribute to have non-smoother (rough) kernel densities which may help to reduce overlapping between classes. In fact, for smooth kernels, probabilities for each class could be very close and do not enable a clear distinction between classes which lead to misclassification. We estimate that, if sufficient number of instances is available for each class, small σ (large N) will contribute to increase accuracy. On contrary, if there are few examples for a class, kernels may be too localized to this class.

As will be seen, such a method contributes to improve the classification accuracy on real datasets as it will be seen in the experimental section. This method exploits a statistical view of the neighborhood since an instance will have a high probability value for a class as soon as its value for each attribute is close to the values of other examples in the class. In the next section, the idea of closeness will be captured by means of a fuzzy set.

3.2 Approximate equality-based interpretations of data

In this section, we propose two other methods for building a possibility distribution directly from a set of data, without computing a Gaussian probability distribution first. This type of approach is well in agreement with the generalized set-like view of possibility distributions, as pointed out in the background section. Indeed, a possibility can take into account the similarity between an observed value of an attribute and other observed values of the same attribute in the training examples. From a logical point of view, one can assume that $\Pi(a_i|c_j) = 1$ as soon as the value a_i has been observed at least one time in association with the class

c_j. Conversely, if a value a'_i has not been observed in association with the class c_j it does not necessarily mean that $\Pi(a'_i|c_j) = 0$. In such case, we may consider that $\Pi(a'_i|c_j)$ should all the closer to 1 as a'_i is closer to an observed value a_i. This non-frequentist idea was first suggested in Dubois and Prade (1993). It is worth emphasizing that this is a purely local view of the building of the distribution, which does not make any assumption on its shape. This type of approach still makes an independent assumption of the attribute with respect to the class.

In this framework, the two suggested classification methods use an *approximate equality relation* between numerical values. Let d be the distance between the two values, this fuzzy relation, namely $\mu_E(d(x, y))$ estimates to what extent x is close to y as follows (in other words E is a fuzzy set with decreasing membership function on [0,1] with a bounded support and such that $\mu_E(0) = 1$):

$$\mu_E(d) = \max\left(0, \min\left(1, \frac{\alpha + \beta - d}{\beta}\right)\right), \quad \alpha \geq 0; \beta > 0. \tag{12}$$

This relation is parameterized by α and β. The parameters α and β are, respectively, fixed to 0 and 1 in (12) for simplicity, once d is normalized in [0, 1], for all attributes. $\alpha = 0$ means that we use a triangular membership function, while $\beta = 1$ means that $\mu_E(d) = 0$ only for the most distant values of attributes. This closeness relation is now used to build a fuzzy histogram from the data.

3.2.1 The Fuzzy Histogram Classifier (FuHC)

Namely, we use the fuzzy relation E to build a fuzzy histogram (Strauss et al. 2000) for attribute a_i given a class c_j, in the following way:

$$\pi(a_i|c_j) = \frac{1}{N_j} \sum_{k=1}^{N_j} \mu_E(d(a_i, a_{ik})), \tag{13}$$

where N_j is the number of instances belonging to the class c_j. The idea is here to be more faithful to the way the data are distributed (rather than assuming a normal distribution) and to take advantage of the approximate equality for obtaining a smooth distribution on the numerical domain, and may be supplying the scarcity of data. In that respect, the parameters of the approximate equality relation, depending on their values, not only reflect the expression of a tolerance on values that are not significantly different for a given attribute, but may also express a form of extrapolation from the observed data. The distribution (13) can then be directly used in the classification procedure. The algorithm based on this method will be named Fuzzy Histogram Classifier (FuHC) in the following. This classifier is a generalized case of a previously proposed classification method for continuous data (Bounhas and Mellouli 2010). In Bounhas and Mellouli (2010), the authors exploited a reduced version of the proximity equality function defined in Eq. (12) and they used a distance metric applied to non-normalized attributes.

3.2.2 Nearest-Neighbor-based Possibilistic Classifier

We propose a second approach, named Nearest-Neighbor-based Possibilistic Classifier (NNPC), which is based only on the analysis of the proximities between the attribute values a_{ik} belonging to each class c_j without counting them. The main idea of this classifier is to search for the nearest neighbor attribute value a_{ik} for the attribute value a_i of the item to be classified, in the training set of each class. The approximate equality function calculated between a_i and its nearest neighbor a_{ik} is then used to estimate the possibility distribution of the attribute value a_i given a class c_j as follows:

$$\pi(a_i|c_j) = \max_{k=1}^{N_j} \mu_E(d(a_i, a_{ik})). \tag{14}$$

In this approach, the closer an attribute value a_i to other attribute values of instances belonging to a class c_j, the greater the possibility to belong to the class (w.r.t. the considered attribute). The expression (12) may be considered as a genuine possibility distribution (Sudkamp 2000). An attribute value having a possibility 0 means that this value is not compatible with the associated class (it is the case when the value is not close to any other observed value of the attribute for the class). If the possibility is equal or close to 1, then the value is relevant for describing the class (a value having a small distance to instances of a class is considered as a possible candidate value in the representation of the class for a considered attribute).

4 Detecting ambiguities in possibilistic classifiers as a basis for improvement

In some cases, classes may have very close plausibility estimates. In the conference version of this work (Bounhas et al. 2010), we have proposed a multiple classification approach to deal with such conflicting situations. Instead of classifying a new instance in the most plausible class, the idea is to consider more than one class at a time when the plausibility difference between the most relevant classes is negligible. Experimental results for the multiple-classification approach showed that, for all datasets, classification accuracy of NPC and NNPC was significantly increased in the case of multiple-classification by comparison with the classical classification. Besides, the accuracy of NBC was not really increased by a similar procedure because the probability rates were generally significantly different. This

is due to the use of product and division applied to numbers that are generally small and to the additive normalization constraint on probabilities. On the basis of these preliminary results, one may expect that possibilistic classifiers will have a better ability to detect confusion between classes than Bayesian ones. In this section, we first discuss how to evaluate ambiguity in classifiers and how to compare possibilistic and Bayesian classifiers in terms of their ability to distinguish between classes. Then, we propose a method to improve the performance of possibilistic classifiers on the basis of the detected ambiguities. These two points will be experimentally validated in the next section.

4.1 Meaningfulness of ambiguity in possibilistic classification

The intuitive idea behind this study is to estimate to what extent classification error is related to the ambiguity between close plausibility evaluations. In order to do that, we first define the classification ambiguity for an instance a with respect to classes as follows:

$$\text{AmbiguityDiff}(a, c_1, \ldots, c_n) = 1 - (\Pi(c_1|a) - \Pi(c_2|a)) \tag{15}$$

where c_1 and c_2 are, respectively, the most and the second most relevant classes for a.

As experimentally checked, such a difference-based ambiguity measure is not appropriate for comparing probability values. Indeed, since these values are obtained as products (and quotient) of usually small values, fixing a threshold that is independent from the data set is not possible in general. For this reason we use a more appropriate ambiguity measure for Bayesian classifiers based on the ratio of probability of the second most relevant class and the first most relevant class:

$$\text{AmbiguityRatio}(a, c_1, \ldots, c_n) = \frac{P(c_2|a)}{P(c_1|a))}. \tag{16}$$

4.2 The hybrid possibilistic classification approach (HPC)

In classification problems, the main issue is to derive a model which is able to predict a unique class for any unseen example. Assigning a unique class to an example in a justified way may become difficult when the available information provided by the training examples is poor. This information may be poor in two respects. First, the training provides only a sampling which may be very scarce in some areas of the attribute space. Besides, the attribute vocabulary may be insufficient for discerning between examples having close descriptions but belonging to different classes. Regardless of the learning method, it may seem difficult to overcome such lacks of information and still justify a unique classification. However, another limitation of the discriminating power of a classifier may come from a systematic independence assumption, as done in naive Bayesian-like classifiers (probabilistic or possibilistic). If we are able to detect when the classification of an example may be problematic, this kind of limitation may be, at least partially, overcome. The idea is to take advantage of the fact that Bayesian-like classifiers allows for an ambiguity analysis based on the plausibility degrees of belonging to a class. Then, problematic classifications may be detected, and in this case, a second algorithm (which does not make the independent assumption) can be applied for breaking the ties.

Thus, we propose to exploit a hybrid possibilistic classification (HPC) approach which aims at improving the accuracy of each of the previously introduced possibilistic classifiers. In this scope, we combine each proposed classifier with a Nearest-Neighbour Heuristic (NNH). The Nearest-Neighbour-based classification is a local method that classifies an example on the basis of its similarity with the training examples in its neighborhood. In our context, NNH has two advantages: (i) its less sensitivity to the violation of the independence assumption, (ii) due to the local nature of NNH an additional computer time cost only occurs in case of ambiguity. We expect that this heuristic may help the Bayesian-like classifiers to choose between classes whose plausibility are too close by preferring one on a nearest-neighbor basis, instead of just choosing the most plausible class, even if the plausibility difference is not significant.

We consider that a classifier is in a failure state if the ambiguity (defined by (15) or (16)) overcomes some fixed threshold ε. Note that, having a too liberal threshold would amount to use only the NNH. The HPC is detailed in the following algorithm:

Algorithm 1 Hybrid Possibilistic Classification Algorithm (PC)

 # PC is a Possibilistic Classifier which can be the NPC, the FNPC, the FuHC or the NNPC
Select an instance a_{ts} to classify
Class$(a_{ts}) \leftarrow \emptyset$
Classify a_{ts} by PC
$c_1 \leftarrow$ Most plausible class
$c_2 \leftarrow$ Second most plausible class
if $Ambiguity(a_{ts}, c_1, c_2) < \varepsilon$ **then**
  Class$(I_{ts}) \leftarrow c_1$
else
  $\pi(Instance1|a_{ts}) \leftarrow NNH(a_{ts}, c_1)$
  $\pi(Instance2|a_{ts}) \leftarrow NNH(a_{ts}, c_2)$
  if $\pi(Instance1|a_{ts}) > \pi(Instance2|a_{ts})$ **then**
    Class$(a_{ts}) \leftarrow c_1$
  else
    Class$(a_{ts}) \leftarrow c_2$
  end if
end if
return Class(a_{ts})

Given an instance a_{ts} to be classified, for each training instance a_{tr} labeled with the class c_j, the NNH estimates the possibility degrees $\pi(a_{tr}|a_{ts})$ as follows:

$$\pi(a_{tr}|a_{ts}) = \pi(a_{1_{(ts)}}|a_{tr}) * \cdots * \pi(a_{M_{(ts)}}|a_{tr}) \quad (17)$$

with

$$\pi(a_{i_{(ts)}}|a_{tr}) = \mu_E(d(a_{i_{(ts)}}, a_{i_{(tr)}})) \quad (18)$$

The * may be the minimum, or the product. We may also think of using the leximin refinement of the minimum that amounts, when comparing two vectors of evaluations, to first reorder them increasingly, and then to reduce the comparison to a minimum-based evaluation of the sub-vectors made of the values where the two reordered vectors are not (approximately) equal. The attribute $a_{i(ts)}$ (respectively, $a_{i(tr)}$) is the attribute of level i of the instance a_{ts} (respectively, a_{tr}).

The Nearest-Neighbour training instance to a_{ts} for the class c_j is the instance having the highest possibility among all instances a_{tr} belonging to the training set labeled with c_j.

$$a^*_{tr} = \arg\max_{a_{tr}} \pi(a_{tr}|a_{ts}) \quad (19)$$

5 Related works

Some approaches have already proposed the use of a possibilistic data representation in classification methods based on decision trees, Bayesian-like or case-based approaches. A general discussion about the appropriateness of fuzzy set methods in machine learning can be found in Hüllermeier (2005). Most of the works in possibilistic classification are motivated by the handling of imprecision and uncertainty about attribute values or the classes. Some assume that there is a partial ignorance about class values. This ignorance, modeled through possibility degrees, reflects the expert knowledge about the possible class of the training instance. In general, the approaches deal with discrete attribute values only and are not appropriate for continuous attributes (and thus require a preliminary discretization phase for the continuous attribute values).

By contrast, the work reported here presents a new type of classification method suitable for training data not pervaded with uncertainty. It relies on a possibilistic representation of the attribute values associated with a class, which offers more flexibility than the classical probabilistic setting. Moreover, we focus on the handling of numerical attributes.

We now provide a brief survey of the literature on possibilistic classification. We start with approaches based on decision trees, before a more detailed discussion on Bayesian classifiers applied to possibilistic data.

Ben Amor et al. (2004) have developed a qualitative approach based on decision trees for classifying examples having uncertain attribute values. Uncertainty on attribute values is represented by means of possibility distributions given by an expert. In Jenhani et al. (2008), possibilistic decision trees are induced from instances associated with categorical attributes and vaguely specified classes. Uncertainty, modeled through possibility theory, concerns only class attribute, whereas other predictive attributes are supposed to be certainly known. The authors developed three approaches for possibilistic decision trees. The first one, using possibilistic distributions in all steps of the tree construction, is based on the non-specificity measure of possibility theory to define an attribute selection measure. The two remaining approaches make use of the notion of similarity between possibility distributions for extending the C4.5 algorithm to support data uncertainty.

A Naive Bayes Style Possibilistic Classifier (NBSPC) is proposed by Borgelt and Gebhardt (1999) to deal with imprecise training sets. For this classifier, imprecision concerns only attribute values of instances (the class attribute and the testing set are supposed to be perfect). Given the class attribute, possibility distributions for attributes are estimated from the computation of the maximum-based projection (Borgelt and Kruse 1988) over the set S of precise instances (S is included in the extended dataset) which contains both the target value of the considered attribute with the class.

A naive possibilistic network classifier, proposed by Haouari et al. (2009), presents a building procedure that deals with imperfect dataset attributes and classes, and a classification procedure used to classify unseen examples which may have imperfect attribute values. This imperfection is modeled through a possibility distribution given by an expert who expresses its partial ignorance, due to a lack of a priori knowledge. There are some similarities between our proposed approach and the one by Haouari et al. (2009). In particular, they are based on the same idea stating that an attribute value is all the more possible if there is an example, in the training set, with the same attribute value (in the discrete case in Haouari et al. 2009) and very close attribute value (in terms of similarity in the numerical case). However, the approach in Haouari et al. (2009) does not require any conditional distribution over attributes to be defined in the certain case, whereas the main focus, in our proposed approaches, is how to estimate such possibility distribution for numerical data in the certain case.

Benferhat and Tabia (2008) propose an efficient algorithm for revising, using Jeffrey's rule, possibilistic knowledge encoded by a naive product-based possibilistic network classifier on the basis of uncertain inputs. The main advantage of the proposed algorithm is its capability

to process the classification task in polynomial time with respected to the number of attributes.

Besides, some case-based classification techniques, which make use of possibility theory and fuzzy sets, are also proposed in the literature. We can particularly mention the possibilistic instance-based learning approach (Hüllermeier 2003). In this work, the author proposes a possibilistic version of the classical instance-based learning paradigm using similarity measures. Interestingly, this approach supports classification and function approximation at the same time. Indeed, the method is based on a general possibilistic extrapolation principle that amounts to state the more similar to a known example the case to be classified is, the more plausible the case and the example should belong to the same class. This idea is further refined in Hüllermeier (2003) by evaluating this plausibility by means of an interval whose lower bound reflects the "guaranteed" possibility of the class, and upper bound the extent to which this class is not impossible. In a more recent work (Beringer and Hüllermeier 2008), the authors develop a bipolar possibilistic method for case-based learning and prediction. This possibilistic instance-based learning approach may look similar to the methods introduced in Sect. 3.2 and to the nearest neighbor heuristics in particular. However, there are differences, although both emphasizes a possibilistic view of classification based on similarity. In Hüllermeier (2003) a conditional possibility of a class given the case description is defined directly, taking into account all the attributes together. In the methods presented in Sect 3.2, we rather start by defining the plausibility of a particular attribute value for a given class (on a similarity basis) and then apply a Bayesian-like machinery for obtaining the classification result. The fact that the similarity idea is applied to attributes one by one makes it necessary to resort to an independence assumption.

6 Experiments and discussion

This section provides experimental results for the possibilistic classifiers that have been previously introduced. The experimental study is based on several datasets taken from the U.C.I repository of machine learning databases (Mertz and Murphy 2000). A brief description of these datasets is given in Table 1. Since we have chosen to deal only with numerical attributes in this study, all these datasets have numerical attribute values.

For each dataset, we used a 10-cross-validation to compare the accuracy of the classifiers. In order to evaluate the accuracy of each classifier, we have used the standard Percent of Correct Classification (PCC) defined as follows:

Table 1 Description of datasets

Database	Data	Attributes	Classes
Iris	150	4	3
W. B. Cancer	699	8	2
Wine	178	13	3
Diabetes	768	7	2
Magic gamma telescope	1074	10	2
Transfusion	748	4	2
Satellite Image	1090	37	6
Segment	1500	20	7
Yeast	1484	9	10
Ecoli	336	8	8
Glass	214	10	7
Iosophere	351	35	2
Letter	3050	17	26
German	1000	25	2
Heart	270	14	2

$$\mathrm{PCC} = \frac{\text{number of well classified instances}}{\text{total number of instances}} * 100 \qquad (20)$$

The experimental study is divided into three parts. First, we evaluate the different possibilistic classifiers NPC, FNPC, FuHC and NNPC methods and we compare our results to those of probabilistic ones, namely NBC (John and Langley 1995) and FNBC (John and Langley 1995). This comparative study is carried out through paired t tests. They are parametric tests that check if the difference between the results of two classifiers over various datasets is significant enough (Demsar 2006). If the null hypothesis (the two compared classifiers have the same accuracy) is rejected, this means that there are statistically significant differences between the two classifiers. The p value gives an idea of how large is this difference. The lower the p value with respect to a threshold (usually 0.05), the more significant the difference between the classifiers.

Note that at this step, we are not handling ambiguity in classification, and then classifiers always assign a class to a considered case. Second, we compare the capabilities of possibilistic and probabilistic classifiers for detecting examples that are ambiguous with respect to classification. Third, we test the ability of the hybrid-classification method for improving the performance of the possibilistic classifiers. We use the signed-ranks test to measure the significance of this improvement.

6.1 Results for the possibilistic classifiers

We use the product in the aggregation step for all possibilistic classifiers, except for the NNPC where we use the minimum because it provides better results for the ambiguity study and it has been shown in Bounhas et al. (2010)

that the three versions (product, minimum, and a leximin-based refinement of minimum) have a competitive efficiency in this case. We only considered normalized attribute values in this paper.

Table 2 shows the classification results obtained with NPC, NBC, FNPC, FNBC, FuHC and NNPC for the 15 mentioned datasets. We also present those of the leximin-based NNH considered here as an independent classifier. By comparing the classification results of the first six classifiers we can notice the following:

- For the two classifiers NPC and NBC, which assume that the attribute values are normally distributed, we remark that NPC is more accurate than NBC on four databases (Yeast, Ecoli, Glass and Heart) and less accurate on the remaining databases except Iris where the two classifiers have the same accuracy. A normality test (test of Shapiro-Wilk) done on these databases (Yeast, Ecoli, Glass and Heart) show that they contain attributes that are not normally distributed. We may suppose that applying a Probability–Possibility transformation on the NBC (which leads to NPC) enables the classifier to be less sensitive to normality violation. As suggested in Sect. 2.3, one may also think that when normality assumption is not supported by the data, especially for datasets with a high number of attributes, the NBC reinforces the error rate (by the use of multiplication), making the NPC more efficient in this case.
- As previously observed in (John and Langley 1995), FNBC is overall better than classical NBC. In fact, FNBC is more accurate than the NBC in seven of the 15 datasets and less accurate in five datasets and not significantly different in three cases (Iris, Diabetes and Satellite Image).
- For the four classifiers using Gaussian distributions (NPC, NBC, FNPC and FNBC), classification results of the FNPC are better than other classifiers for all datasets except in the case of "Transfusion" and "Yeast" databases where FNPC performs worse than others.
- If we compare results for the two flexible classifiers (FNPC and FNBC), we note that the FNPC performs better with the highest accuracy for the majority of datasets. For this classifier, the greatest increase in accuracy compared with the FNBC has occurred for databases "Glass", "Ecoli", "Satellite image", "Segment" and "Letter" (Table 2). In Table 1, we note that the attributes for these databases range from 8 to 37, and the number of classes from 6 to 26. So the FNPC is significantly more efficient than FNBC (and also than NPC and NBC) for datasets with a high number of attributes and classes.

Table 2 Experimental results given as the mean and the standard deviation of 10 cross-validations

	NPC	NBC	FNPC	FNBC	FuHC	NNPC	NNH
Iris	95.33 ± 6.0 (4)	95.33 ± 6.0 (4)	**96.0 ± 5.33** (2)	95.33 ± 5.21 (4)	94.66 ± 4.0 (6)	90.66 ± 4.42 (7)	**96.0 ± 4.42** (2)
Cancer	95.03 ± 2.26 (6)	96.34 ± 0.97 (3)	97.37 ± 1.82 (2)	**97.65 ± 1.76** (1)	96.05 ± 1.96 (5)	93.41 ± 2.49 (7)	96.06 ± 1.82 (4)
Wine	94.37 ± 5.56 (4)	**97.15 ± 2.86** (1)	96.6 ± 3.73 (2.5)	96.67 ± 5.67 (2.5)	93.26 ± 4.14 (5.5)	92.64 ± 5.12 (7)	93.26 ± 5.98 (5.5)
Diabetes	69.01 ± 3.99 (5)	74.34 ± 4.44 (3)	**74.36 ± 4.57** (1)	74.35 ± 3.38 (2)	73.44 ± 5.31 (4)	67.96 ± 6.05 (7)	67.97 ± 5.73 (6)
Magic	59.24 ± 7.09 (7)	66.02 ± 5.37 (5)	73.37 ± 2.96 (2)	72.8 ± 3.29 (3)	68.34 ± 6.69 (4)	64.80 ± 2.41 (6)	**74.21 ± 4.51** (1)
Transfusion	61.67 ± 6.6 (7)	72.6 ± 4.56 (3)	67.43 ± 7.43 (6)	70.09 ± 7.68 (4)	72.76 ± 7.19 (2)	**76.50 ± 5.94** (1)	68.73 ± 5.61 (5)
Sat. Image	88.26 ± 2.62 (6)	90.55 ± 2.46 (4)	92.02 ± 2.81 (3)	90.0 ± 4.39 (5)	86.88 ± 3.67 (7)	93.58 ± 1.88 (2)	**93.95 ± 2.6** (1)
Segment	71.47 ± 4.15 (7)	80.73 ± 2.16 (6)	90.73 ± 1.8 (2.5)	88.27 ± 3.19 (4)	81.07 ± 3.51 (5)	90.73 ± 2.15 (2.5)	**95.07 ± 1.61 (1)**
Yeast	49.67 ± 4.87 (5)	48.65 ± 4.42 (6)	52.02 ± 5.05 (4)	**55.93 ± 3.36** (1)	53.36 ± 4.57 (2)	43.06 ± 2.53 (7)	52.16 ± 3.47 (3)
Ecoli	83.37 ± 4.46 (2)	82.53 ± 5.32 (3)	**83.55 ± 9.4 (1)**	79.02 ± 10.0 (6)	77.7 ± 13.31 (7)	80.65 ± 6.98 (4)	79.39 ± 9.22 (5)
Glass	49.18 ± 11.8 (5)	33.74 ± 9.0 (7)	58.46 ± 9.59 (3)	53.42 ± 16.0 (4)	39.26 ± 13.9 (6)	65.84 ± 9.70 (2)	**67.93 ± 7.65 (1)**
Isosphere	58.4 ± 10.95 (7)	69.23 ± 7.85 (6)	**91.75 ± 4.11 (1)**	90.88 ± 4.0 (3)	79.77 ± 9.6 (5)	91.45 ± 4.24 (2)	88.33 ± 3.87 (4)
Letter	60.42 ± 3.24 (5)	63.28 ± 2.13 (3)	72.3 ± 2.87 (2)	61.61 ± 1.97 (4)	50.36 ± 2.33 (6)	35.47 ± 3.2 (7)	**82.56 ± 1.92 (1)**
German	66.4 ± 3.97 (7)	68.5 ± 3.29 (5)	**71.8 ± 4.21 (1)**	70.0 ± 4.96 (2)	69.1 ± 2.88 (4)	69.8 ± 5.47 (3)	66.6 ± 3.26 (6)
Heart	**84.08 ± 8.77 (1)**	83.7 ± 6.87 (2)	83.33 ± 9.55 (3)	82.96 ± 7.8 (4)	81.11 ± 8.19 (5)	58.89 ± 7.49 (7)	71.11 ± 8.73 (6)
Average rank	5.2	4.06	**2.4**	3.3	4.9	4.7	3.4

Bold values reflect the best classifier in terms of accuracy

- Experiments of the second family made of the approximate equality-based classifiers (FuHC, NNPC and NNH) show that they have a competitive efficiency with respect to other possibilistic classifiers for the majority of databases. Besides, we note that the leximin-based NNH not only outperforms the FuHC and also the NNPC, but also all other classifiers for 5 datasets (Magic, Satellite Image, Segment, Glass and Letter). Table 1 shows that these datasets have a higher number of attributes, classes and instances. Thus, the leximin-based NNH seems to be the most efficient classifier for datasets with high dimensionality. Indeed, in contrast to the product-based evaluation, the leximin evaluation is not very sensitive to the dimension of the attributes universe and then the methods based on this evaluation may be expected to be more robust.

The average ranks given between parentheses in Table 2 confirm what we have already noted above. On average, the FNPC ranks the first (with rank 2.4) while the FNBC and the NNH rank the second (respectively, 3.3 and 3.4).

Figure 2 shows the results of the paired t test between the FNPC and all the other classifiers, whereas Fig. 3 shows results between the NNH and all other classifiers. We choose to compare the two best-ranked possibilistic classifiers with others for a deeper comparison. Dots above the abscissa axes in Fig. 2 (respectively, Fig. 3) reflect datasets for which the FNPC (respectively, the NNH) is significantly better than the compared classifier. Dots under the abscissa axes reflect datasets for which the FNPC (or the NNH) is significantly worse than the second classifier. For the datasets where the two classifiers have an equivalent accuracy ($p > 0.05$), dots are on the abscissa axes. The

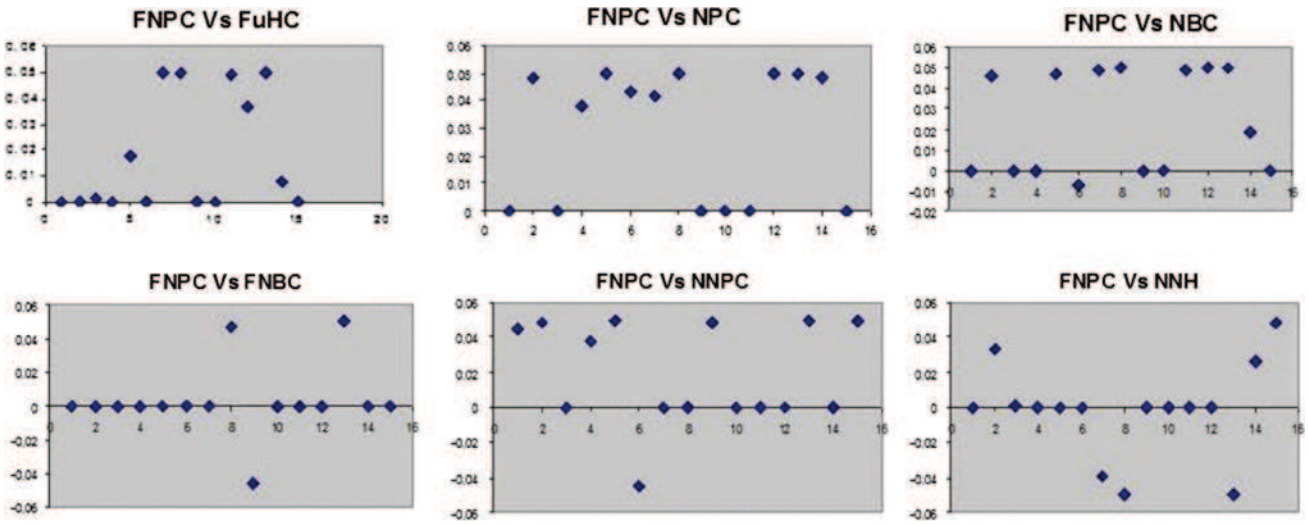


Fig. 2 Results of the paired t test between the FNPC and other classifiers

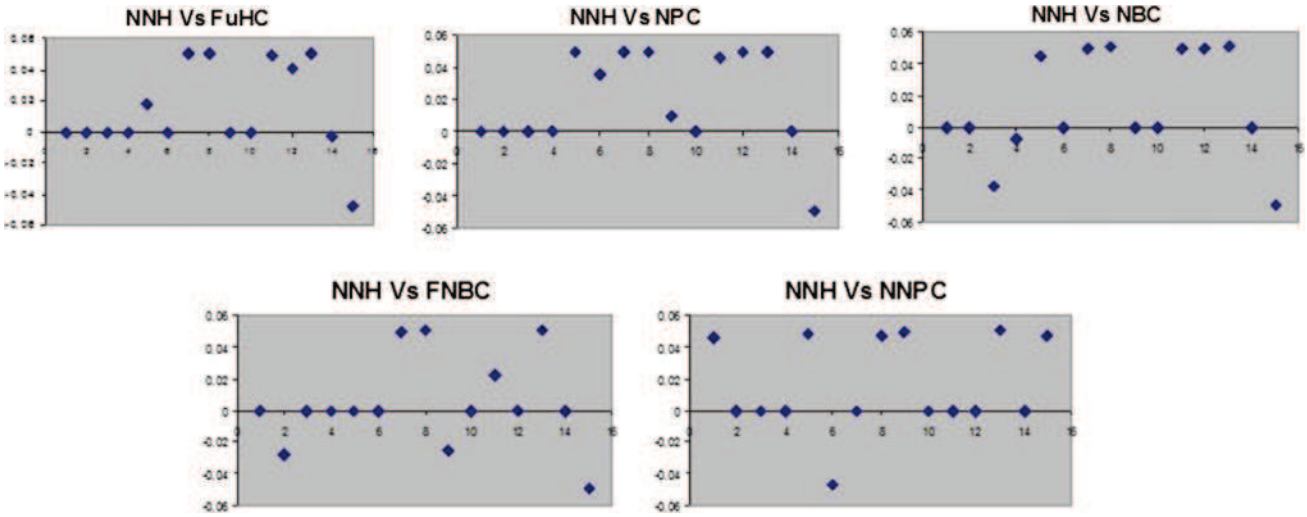


Fig. 3 Results of the paired t test between the NNH and other classifiers

datasets in these comparisons are considered in the same order as in Table 1.

Results of the paired t test shows that the proposed FNPC *significantly outperforms* the FuHC, NPC, NBC, and the NNPC in terms of the number of datasets where the FNPC has a significantly better accuracy than the compared classifier. We can also see in Fig. 2 that the FNPC is slightly more accurate than the FNBC (because it is significantly more accurate than the latter in two datasets and less accurate in only one dataset) and is equivalent in terms of accuracy to the NNH (it is significantly better in three datasets, worst in three others and equivalent in the remaining datasets).

By comparing the NNH with the other classifiers, we observe similar results as for the FNPC. In fact, the paired t test in Fig. 3 proves that the NNH is *much better* than any other classifier except for the FNBC where the NNH is better on four datasets, worst on three datasets and equivalent on the remaining.

These results show that the FNPC and the NNH are the most efficient possibilistic classifiers among the proposed ones, and they at least compete with the classical and the flexible Bayesian classifiers. Especially, they are slightly better for datasets with a large number of attributes, classes and instances.

6.2 Results of the ambiguity study between near classes

As explained in Sect. 4.1, we are interested in a possible relationship between classification ambiguity and classification errors in the case of possibilistic and Bayesian classifiers.

For each classifier, we fix n levels ($n = 5$ in this experiment) of ambiguity using n intervals having the same length that partition the interval [0, 1]. Then for each ambiguity interval, we compute the number of correctly classified examples (CCE) and the number of incorrectly classified ones (ICE) in the testing set. Experimental results for the NNPC, NPC and the NBC are given, respectively, in Figs. 4, 5 and 6. In each figure, we present the amount of ICE and CCE for each classifier for datasets "Segment" and "Sat-Image" (part a and c). We also exhibit the frequency of error calculated by the ratio: ICE/(CCE + ICE) for the two datasets in part b and d in each figure. Figure 7 summarises results of the error frequency comparison between the three studied classifiers.

We note that ambiguity levels (AL_i in Figs. 4, 5 and 6) represent the n intervals of the possibility–probability difference between the most relevant classes ranging in [0, 1] and they are chosen in a decreasing manner such that AL_1 corresponds to the highest ambiguity level, whereas AL_n corresponds to the lowest ambiguity level. Results given in Figs. 4, 5 and 6 for the CCE and the ICE are an averaged number though the 10-cross-validations for the NNPC, NPC and NBC.

In Figs. 4 and 5, we can see that the frequency of incorrectly classified instances (part b and d) decreases when the ambiguity decreases. These figures illustrate also that the highest frequency of incorrect classified instances corresponds to the case of the first ambiguity level that

Fig. 4 Results for the NNPC

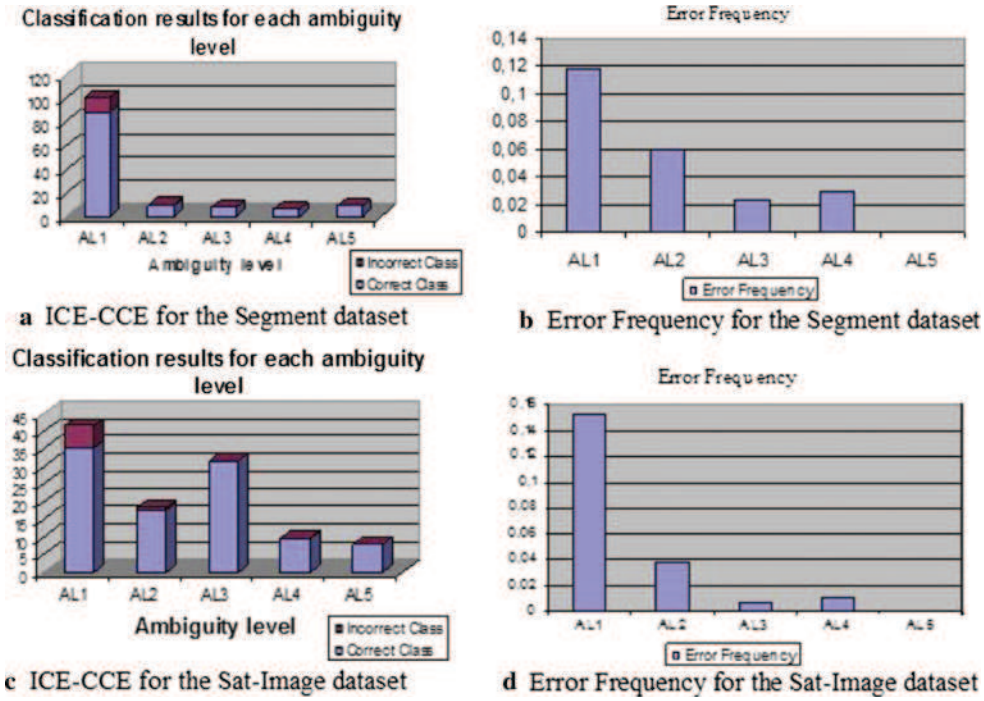
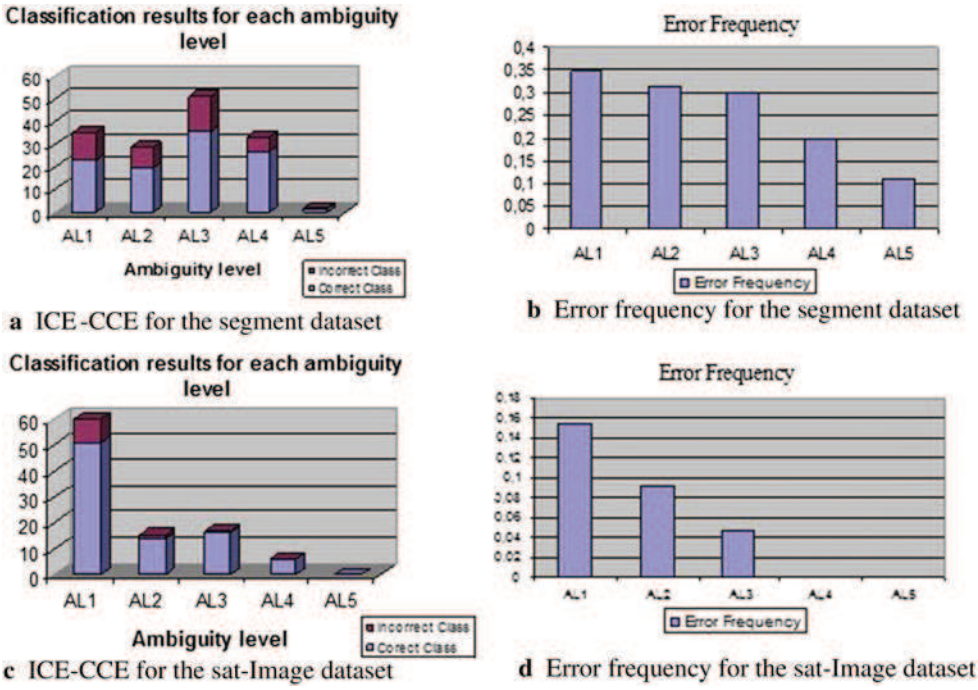


Fig. 5 Results for the NPC



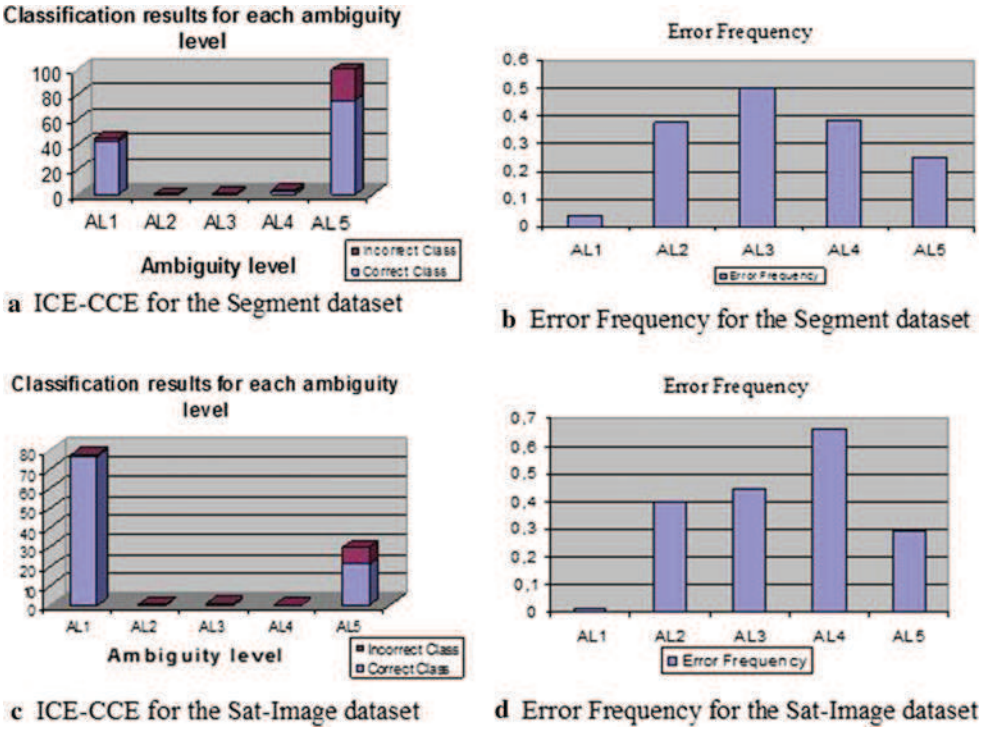
a ICE-CCE for the segment dataset

b Error frequency for the segment dataset

c ICE-CCE for the sat-Image dataset

d Error frequency for the sat-Image dataset

Fig. 6 Results for the NBC



a ICE-CCE for the Segment dataset

b Error Frequency for the Segment dataset

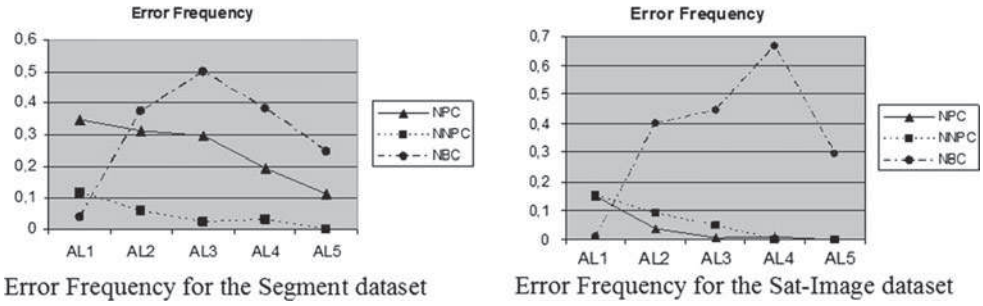
c ICE-CCE for the Sat-Image dataset

d Error Frequency for the Sat-Image dataset

reflects the highest ambiguity. We also notice that, for the lowest ambiguity level (AL_4 and AL_5), possibilistic classifiers make almost no error (ICE $\approx$ 0 even if CCE is always relatively high). These results are nearly the same for the two classifiers NNPC and NPC for almost all datasets. Here we keep only the "Segment" and "Sat-Image" as an illustrative example.

From Figs. 4 and 5, we can see that the higher the ambiguity, the greater the error rate is and the lower the ambiguity is, the more the classifier is able to detect the correct class. So we can say that there is a relationship between ambiguity and classification accuracy for possibilistic classifiers. These results are clearly confirmed by the results shown in Fig. 7.

Fig. 7 Error frequency for the three classifiers



However, in Fig. 6 (and also in Fig. 7), corresponding to the case of NBC, we note that the frequency of error has a non-steady behavior. For the two datasets Segment and sat-Image, instances are either classified with a high ambiguity in AL_1, or much discriminated in AL_5. Moreover, the error rate for this classifier seems to be greater for the lowest ambiguity level than that for the highest one. The error frequency remains higher than 30 % for the lowest ambiguity level. So, we can say that in spite of the fact that the NBC distinguishes well between classes in AL_5, it makes more errors in classification. This means that the high distinction ability between classes in this case has no particular meaning and may be simply caused by the exponential nature of Gaussian densities.

These results support the intuition underlying the use of possibilistic classifiers. In fact, this study shows that these classifiers are able to detect conflicts in case of ambiguous classification and to acknowledge difficulties in classifying a conflicting instance. On the contrary, Bayesian Classifiers, due to the difficulty to have a faithful and general measure of ambiguity, seem to have a lower capability for detecting such conflicting situations.

6.3 Results of the hybrid possibilistic classification

Table 3 includes experimental results for NPC, NBC, FuHC and NNPC in the case of hybrid possibilistic classification.

In this case, we use the Nearest-Neighbor Heuristic to help classifying a new instance (instead of only considering the main classifier), when classes have very close plausibility evaluations, i.e., if the difference between their plausibility is less than a fixed level. In our experimental study, this level is fixed to 0.1 (i.e. ambiguity level greater than 0.9) for all classifiers. We choose a relatively high threshold to show the effect of the hybrid classification for all possibilistic classifiers at the same time. In fact, the FNPC distinguishes well between classes when compared with NPC or FuHC (the difference between class possibilities is relatively high), so with a low threshold, the hybrid classification would not have any effect on the classical FNPC.

We evaluate the effect of the hybrid classification and its ability to improve the accuracy of possibilistic classifiers. For each classifier, we compare the classification accuracy

Table 3 Experimental results for the hybrid possibilistic classification

	NPC+NNH	FNPC+NNH	FuHC+NNH	NNPC+NNH	NNH
Iris	96.67 ± 4.47	96.67 ± 6.15	96.66 ± 3.34	96.0 ± 6.11	96.0 ± 4.42
Cancer	95.18 ± 1.83	97.36 ± 2.85	96.35 ± 2.27	95.76 ± 3.1	96.06 ± 1.82
Wine	94.93 ± 4.63	97.22 ± 3.73	93.19 ± 4.32	93.89 ± 3.89	93.26 ± 5.98
Diabetes	71.49 ± 4.66	74.1 ± 5.42	69.03 ± 4.29	68.21 ± 5.32	67.97 ± 5.73
Magic	65.46 ± 6.73	74.95 ± 3.23	73.37 ± 4.92	72.72 ± 3.13	74.21 ± 4.51
Transfusion	65.78 ± 6.11	72.22 ± 5.81	71.02 ± 4.35	72.33 ± 2.97	68.73 ± 5.61
Sat. Image	88.53 ± 4.94	92.57 ± 2.48	92.48 ± 1.35	95.05 ± 1.55	93.95 ± 2.6
Segment	75.67 ± 3.02	92.93 ± 2.31	91.73 ± 1.91	95.6 ± 2.09	95.07 ± 1.61
Yeast	54.78 ± 2.83	54.99 ± 3.34	54.51 ± 3.24	48.78 ± 2.02	52.16 ± 3.47
Ecoli	84.26 ± 5.5	84.47 ± 5.54	81.14 ± 8.71	80.47 ± 6.01	79.39 ± 9.22
Glass	59.66 ± 9.75	68.42 ± 9.68	50.0 ± 10.79	66.34 ± 5.42	67.93 ± 7.65
Iosophere	62.71 ± 6.22	92.3 ± 3.15	86.6 ± 7.13	88.34 ± 5.55	88.33 ± 3.87
Letter	68.29 ± 3.14	76.95 ± 2.42	67.1 ± 5.41	50.79 ± 2.96	82.56 ± 1.92
German	69.20 ± 3.12	68.7 ± 3.41	68.3 ± 3.66	67.4 ± 4.39	66.60 ± 3.26
Heart	82.96 ± 7.98	81.85 ± 6.3	78.15 ± 8.19	71.85 ± 6.02	71.11 ± 8.73

Table 4 Results for the Wilcoxon Matched-Pairs Signed-Ranks Test

NPC versus (NPC + NNH)	FNPC versus (FNPC + NNH)	FuHC versus (FuHC + NNH)	NNPC versus (NNPC + NNH)
$p \leq 0.002441$	$p \leq 0.1016$	$p \leq 0.03271$	$p \leq 0.04187$

with or without applying the NNH. For example, in the case of the NPC, we compare column 1 in Table 2 with column 1 in Table 3.

We are only interested here in knowing if the hybrid classification method improves the initial classifier. For doing this, we use the Wilcoxon Matched-Pairs Signed-Ranks Test as proposed by Demsar (2006) since it allows for a direct comparison of the methods without resorting to an analysis of the results on each data set (as done with the paired *t* test). It is a non-parametric alternative to the paired *t* test that enables us to compare two classifiers (or two versions of the same classifier) over multiple datasets. The Signed-Ranks Test ranks the differences in accuracy for each dataset without regard to the sign of the difference and compares the ranks for the positive and the negative differences.

Table 4 includes the *p* values for the comparison of each classical possibilistic classifier with its hybrid version where we combine this classifier with the NNH.

Results in Table 4 show that the hybrid classification contributes to significantly improve the accuracy of the NPC, the FuHC and the NNPC ($p < 0.05$). Although there is an improvement of accuracy in the case of the FNPC for some datasets (Transfusion, Segment, Yeast, Glass, and Letter), this improvement is not statistically significant for all datasets ($p \leq 0.1016$). By comparing the accuracy of the hybrid version of FNPC with the classical FNBC over the 15 datasets, we note that the FNPC + NNH is better than the FNBC with a $p \leq 0.00488$ (instead of a $p \leq 0.05225$ when comparing classical FNPC with FNBC).

These results are not surprising since we have already seen in the first experimental study that the NNH is better than the NPC, FuHC and the NNPC and it is equivalent in terms of accuracy to the FNPC. So we can conclude that combining the NNH with possibilistic classifiers in the hybrid approach contributes only to significantly improve the accuracy of classifiers with lower performance than that of the NNH. However, the hybrid classification does not contribute to significantly improve the performance of the FNPC because the NNH and the FNPC have almost the same classification performance.

7 Conclusion and discussion

The possibilistic classifiers (Haouari et al. 2009; Jenhani et al. 2008) that have been proposed until now are only suitable for discrete attributes. This work has investigated a possibilistic classification paradigm that may be viewed as a counterpart of Bayesian classification and that applies to continuous attribute domains. Then an important issue is the estimation of possibilistic distributions from numerical data, without discretization. For this purpose, we have proposed and tested the performance of two families of possibilistic classifiers: the first family, called Gaussian-based Possibilistic Classifiers, assumes normality assumption when estimating possibilistic distributions. For this family of classifiers, we have used a probability–possibility transformation method enabling us to derive a possibilistic distribution from a probabilistic one. First, we have applied the transformation method to move from a classical NBC to a NPC, which introduces some further tolerance in the description of classes. Then, we have tested the feasibility of a Flexible Naive Possibilistic Classifier, which is the possibilistic counterpart of the Flexible Naive Bayesian Classifier. The FNPC estimates possibilistic distributions in a non-parametric way by applying the transformation method to kernel densities instead of Gaussian ones. The intuition behind this classifier is that kernel densities are less sensible than Gaussian ones to normality violation.

The second family of possibilistic classifiers abandons the normality assumption and has a direct representation of data. We have proposed two other classifiers named Fuzzy Histogram Classifier and Nearest Neighbor-based Possibilistic Classifier in this context. The two proposed classifiers exploit an idea of proximity between attribute values to estimate possibility distributions. In the first classifier, one computes an average proximity, whereas for the second one we analyse proximities between attributes without counting them. The main advantage of this family of classifiers, when compared to the first one, is their ability to derive possibilistic distributions without the need of a normality assumption, which may lead to a more realistic representation of data. Moreover, we have shown that possibilistic classifiers have a higher ability to detect ambiguity between classes than Bayesian classifiers. Namely possibilistic classifiers acknowledge the fact that it is difficult to classify some examples by assessing close possibility degrees to competing classes. In the same situation, Bayesian classifiers may give the illusion of discriminating between classes by assessing very different probability degrees to them.

As an attempt to improve the performance of possibilistic classifiers, we have proposed an hybrid classification method that is based on a Nearest-Neighbor Heuristic used for separating classes having close plausibility estimates. The Nearest-Neighbor Heuristic contributes to help the main classifier to converge to the correct class label in case data information is insufficient for a more precise classification, rather than choosing between classes having very close plausibility estimates in a rather arbitrary way.

We have tested the proposed possibilistic classifiers on several datasets from the UCI repository. Experimental results show the performance of these classifiers for handling numerical input data. However, while the NPC is less sensible than NBC to normality violation, the FNPC shows high classification accuracy and good ability to deal with any type of data when compared with other classifiers in the same family. On the other hand, possibilistic classifiers exploiting proximity between attribute values are competitive with others. Besides, the NNH seems to be the most efficient classifier in particular for databases with high dimensionality. The hybrid classification method exhibits an improvement of the accuracy of possibilistic classifiers, in particular those having a great confusion level between classes which produce close plausibility estimates for classes, such as the NPC. Future research includes the extension of possibilistic classifiers to handle uncertainty in data representation and to deal with imprecise/uncertain attributes and classes; see Bounhas et al. (2011) for preliminary results on these last issues.

Appendix: Naive Bayesian Classifiers

Naive Bayesian Classifiers (NBC) are based on Bayes rule. They assume the independence of the input variables. Despite their simplicity, NBC can often outperform more sophisticated classification methods (Langley et al. 1992). A NBC can be seen as a Bayesian network in which predictive attributes are assumed to be conditionally independent given the class attribute.

Given a vector $X = \{x_1, x_2, \ldots, x_n\}$ to be classified, a NBC computes the posterior probability $P(c_j|X)$ for each class c_j in a set of possible classes $C = (c_1, c_2, \ldots, c_m)$ and labels the case X with the class c_j that achieves the highest posterior probability, that is:

$$c^* = \arg\max_{c_j} P(c_j|X). \tag{21}$$

Using the Bayes rule:

$$P(c_j|x_1, x_2, \ldots, x_n) = \frac{P(x_1, x_2, \ldots, x_n|c_j) * P(c_j)}{P(x_1, x_2, \ldots, x_n)} \tag{22}$$

The denominator $P(x_1, x_2, , x_n)$ is a normalizing factor that can be ignored when determining the maximum *posterior* probability of a class, as it does not depend on the class. The key term in Eq. (2) is $P(x_1, x_2, , x_n|c_j)$ which is estimated from training data. Since Naive Bayes assumes that conditional probabilities of attributes are statistically independent we can decompose the likelihood into a product of terms:

$$P(x_1, x_2, \ldots, x_n|c_j) = \prod_{i=1}^{n} p(x_i|c_j) \tag{23}$$

Even under the independence assumption, the NBC have shown good performance for datasets containing dependent attributes. Domingos and Pazzani (2002) explain that attribute dependency does not strongly affect the classification accuracy. They also relate good performance of NBC to the zero-one loss function which considers that a classifier is successful when the maximum probability is assigned to the correct class (even if estimated probability is inaccurate). The work in Zhang (2004) gives a deeper explanation about the reasons for which the efficiency of NBC is not affected by attribute dependency. The author shows that, even if attributes are strongly dependent (if we look at each pairs of attributes), the global dependencies among all attributes could be insignificant because dependencies may cancel each other out and so they do not affect classification.

The most well-known Bayesian classification approach uses an estimation based on a multinomial distribution over the discretized variables and leads to so-called multinomial classifiers. Such a classifier, which handles only discrete attributes (continuous attributes must be discretized), assumes that all attributes follow a multinomial probability distribution. A variety of multinomial classifiers have been proposed for handling an arbitrary number of independent attributes. Let us mention especially (Langley et al. 1992; Langley and Sage 1994; Grossman and Dominigos 2004), semi-naive Bayesian classifiers (Kononenko 1991; Denton and Perrizo 2004), tree-augmented naive Bayesian classifiers (Friedman et al. 1997), k-dependence Bayesian classifiers (Sahami 1996) and Bayesian Network-augmented naive Bayesian classifiers (Cheng and Greiner 1999).

A second family of NBC is suitable for continuous attribute values. They directly estimate the true density of attributes using *parametric* density. A supplementary common assumption made by the NBC in that case is that within each class the values of numeric attributes are normally distributed around the mean, and they model each attribute through a single Gaussian. Then, the NBC represent such a distribution in terms of its *mean* and *standard deviation* and compute the probability of an observed value from such estimates. This probability is calculated as follows:

$$p(x_i|c_j) = g(x_i, \mu_j, \sigma_j) = \frac{1}{\sqrt{2\Pi}\sigma_j} e^{-\frac{(x_i-\mu_j)^2}{2\sigma_j^2}} \quad (24)$$

The Gaussian classifiers (Geiger and Heckerman 1994; John and Langley 1995) are known for their simplicity and have a smaller complexity, compared with other non-parametric approximations. Although the normality assumption may be a valuable approximation for many benchmarks, it is not always the best estimation. Moreover, if the normality assumption is violated, classification results of NBC may deteriorate.

Other approaches using a non-parametric estimation are those breaking with the strong parametric assumption. The main approaches are based on the mixture model (Figueiredo and Leit ao 1999; McLachlan and Peel 2000) and the Gaussian mixture models (Bishop 1999; McLachlan and Peel 2000). Other approaches use kernel densities (John and Langley 1995; Pérez et al. 2009), leading to so-called Flexible Classifiers. This name is due to the ability of such classifier to represent densities with more than one mode in contrast with simple Gaussian classifiers. Flexible classifiers represent densities of different shapes with high accuracy; however, it results into a considerable increase in complexity.

John and Langley (1995) have proposed a Flexible Naive Bayesian Classifier (FNBC) that abandons the normality assumption and instead uses nonparametric kernel density estimation for each conditional distribution. The FNBC has the same properties as those introduced for the NBC; the only difference is instead of estimating the density for each continuous attribute x by a single Gaussian $g(x, \mu_j, \sigma_j)$, this density is estimated using an averaged large set of Gaussian kernels. To compute continuous attribute density for a specific class j, FNBC calculates n Gaussians, where each of them stores each attribute value encountered during training for this class and then takes the average of the n Gaussians to estimate $p(x_i|c_j)$. More formally, probability distribution is estimated as follows:

$$p(x_i|c_j) = \frac{1}{N_j} \sum_{k=1}^{N_j} g(x_i, \mu_{ik}, \sigma_j) \quad (25)$$

where k ranges over the training set of attribute x_i in class c_j, N_j is the number of instances belonging to the class c_j. The mean μ_{ik} is equal to the real value of attribute i of the instance k belonging to the class j, e.g. $\mu_{ik} = x_{ik}$. For each class j, FNBC estimates this standard deviation by

$$\sigma_j = \frac{1}{\sqrt{N_j}} \quad (26)$$

The authors also prove kernel estimation consistency using (26) (see John and Langley 1995, for details). It has been shown that the kernel density estimation used in the FNBC and applied on several datasets enables this classifier to perform well in datasets where the parametric assumption is violated with little cost for datasets where it holds.

Pérez et al. (2009) have recently proposed a new approach for Flexible Bayesian classifiers based on kernel density estimation that extends the FNBC proposed by John and Langley (1995) to handle dependent attributes and abandons the independence assumption. In this work, three classifiers: tree-augmented naive Bays, a k-dependence Bayesian classifier and a complete graph are adapted to the support kernel Bayesian network paradigm.

References

Ben Amor N, Mellouli K, Benferhat S, Dubois D, Prade H (2002) A theoretical framework for possibilistic independence in a weakly ordered setting. Int J Uncertain Fuzziness Knowledge-Based Syst 10:117–155

Ben Amor N, Benferhat S, Elouedi Z (2004) Qualitative classification and evaluation in possibilistic decision trees. In: FUZZ-IEEE'04, vol 1, pp 653–657

Benferhat S, Tabia K (2008) An efficient algorithm for naive possibilistic classifiers with uncertain inputs. In: Proceedings of 2nd international conference on scalable uncertainty management (SUM'08). LNAI, vol 5291. Springer, Berlin, pp 63–77

Beringer J, Hüllermeier E (2008) Case-based learning in a bipolar possibilistic framework. Int J Intell Syst 23:1119–1134

Bishop CM (1996) Neural networks for pattern recognition. Oxford University Press, New York

Bishop CM (1999) Latent variable models. In: Learning in graphical models, pp 371–403

Borgelt C, Gebhardt J (1999) A naïve bayes style possibilistic classifier. In: Proceedings of 7th European congress on intelligent techniques and soft computing, pp 556–565

Borgelt C, Kruse R (1988) Efficient maximum projection of database-induced multivariate possibility distributions. In: Proceedings of 7th IEEE international conference on fuzzy systems, pp 663–668

Bounhas M, Mellouli K (2010) A possibilistic classification approach to handle continuous data. In: Proceedings of the eighth ACS/IEEE international conference on computer systems and applications (AICCSA-10), pp 1–8

Bounhas M, Mellouli K, Prade H, Serrurier M (2010) From bayesian classifiers to possibilistic classifiers for numerical data. In: Proceedings of the fourth international conference on scalable uncertainty management, pp 112–125

Bounhas M, Prade H, Serrurier M, Mellouli K (2011) Possibilistic classifiers for uncertain numerical data. In: Proceedings of 11th European conference on symbolic and quantitative approaches to reasoning with uncertainty (ECSQARU'11), Belfast, UK, June 29–July 1. LNCS, vol 6717. Springer, Berlin, pp 434–446

Cheng J, Greiner R (1999) Comparing bayesian network classifiers. In: Proceedings of the 15th conference on uncertainty in artificial intelligence, pp 101–107

Cover TM, Hart PE (1967) Nearest neighbour pattern classification. IEEE Trans Inf Theory 13:21–27

De Cooman G (1997) Possibility theory. Part I: measure- and integral-theoretic ground- work. Part II: conditional possibility; Part III: possibilistic independence. Int J Gen Syst 25:291–371

Demsar J (2006) Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 7:1–30

Denton A, Perrizo W (2004) A kernel-based semi-naive Bayesian classifier using p-trees. In: Proceedings of the 4th SIAM international conference on data mining

Devroye L (1983) The equivalence of weak, strong, and complete convergence in l1 for kernel density estimates. Ann Stat 11:896–904

Domingos P, Pazzani M (2002) Beyond independence: conditions for the optimality of the simple bayesian classifier. Mach Learn 29:102–130

Dubois D (2006) Possibility theory and statistical reasoning. Comput Stat Data Anal 51:47–69

Dubois D, Prade H (1988) Possibility theory: an approach to computerized processing of uncertainty

Dubois D, Prade H (1990) Aggregation of possibility measures. In: Multiperson decision making using fuzzy sets and possibility theory, pp 55–63

Dubois D, Prade H (1990) The logical view of conditioning and its application to possibility and evidence theories. Int J Approx Reason 4:23–46

Dubois D, Prade H (1992) When upper probabilities are possibility measures. Fuzzy Sets Syst 49:65–74

Dubois D, Prade H (1993) On data summarization with fuzzy sets. In: Proceedings of the 5th international fuzzy systems association. World Congress (IFSA'93)

Dubois D, Prade H (1998) Possibility theory: qualitative and quantitative aspects. In: Gabbay D, Smets P (eds) Handbook on defeasible reasoning and uncertainty management systems, vol 1, pp 169–226

Dubois D, Prade H (2000) An overview of ordinal and numerical approaches to causal diagnostic problem solving. In: Gabbay DM, Kruse R (eds) Abductive reasoning and learning, handbooks of defeasible reasoning and uncertainty management systems, drums handbooks, vol 4, pp 231–280

Dubois D, Prade H (2009) Formal representations of uncertainty. In: Bouyssou D, Dubois D, Pirlot M, Prade H (eds) Decision-making—concepts and methods, pp 85–156

Dubois D, Prade H, Sandri S (1993) On possibility/probability transformations. Fuzzy Logic, pp 103–112

Dubois D, Laurent F, Gilles M, Prade H (2004) Probability-possibility transformations, triangular fuzzy sets, and probabilistic inequalities. Reliable Comput 10:273–297

Figueiredo M, Leitao JMN (1999) On fitting mixture models. In: Energy minimization methods in computer vision and pattern recognition, vol 1654, pp 732–749

Friedman N, Geiger D, Goldszmidt M (1997) Bayesian network classifiers. Mach Learn 29:131–161

Geiger D, Heckerman D. (1994) Learning gaussian networks. Technical report, Microsoft Research, Advanced Technology Division

Grossman D, Dominigos P (2004) Learning Bayesian maximizing conditional likelihood. In: Proceedings on machine learning, pp 46–57

Haouari B, Ben Amor N, Elouadi Z, Mellouli K (2009) Naive possibilistic network classifiers. Fuzzy Sets Syst 160(22):3224–3238

Hüllermeier E (2003) Possibilistic instance-based learning. Artif Intell 148(1–2):335–383

Hüllermeier E (2005) Fuzzy methods in machine learning and data mining: status and prospects. Fuzzy Sets Syst 156(3):387–406

Jenhani I, Ben Amor N, Elouedi Z (2008) Decision trees as possibilistic classifiers. Int J Approx Reason 48(3):784–807

John GH, Langley P (1995) Estimating continuous distributions in Bayesian classifiers. In: Proceedings of the 11th conference on uncertainty in artificial intelligence

Kononenko I (1991) Semi-naive bayesian classifier. In: Proceedings of the European working session on machine learning, pp 206–219

Kotsiantis SB (2007) Supervised machine learning: a review of classification techniques. Informatica 31:249–268

Langley P, Sage S (1994) Induction of selective bayesian classifiers. In: Proceedings of 10th conference on uncertainty in artificial intelligence (UAI-94), pp 399–406

Langley P, Iba W, Thompson K (1992) An analysis of bayesian classifiers. In: Proceedings of AAAI-92, vol 7, pp 223–228

McLachlan GJ, Peel D (2000) Finite mixture models. Probability and mathematical statistics. Wiley, New York

Mertz J, Murphy PM (2000) Uci repository of machine learning databases. ftp://ftp.ics.uci.edu/pub/machine-learning-databases

Pearl J (1988) Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmman, San Francisco

Pérez A, Larraoaga P, Inza I (2009) Bayesian classifiers based on kernel density estimation: flexible classifiers. Int J Approx Reason 50:341–362

Quinlan JR (1986) Induction of decision trees. Mach Learn 1:81–106

Sahami M (1996) Learning limited dependence bayesian classifiers. In: Proceedings of the 2nd international conference on knowledge discovery and data mining, pp 335–338

Shafer G (1976) A mathematical theory of evidence. Princeton University Press, Princeton

Solomonoff R (1964) A formal theory of inductive inference. Inf Control 7:224–254

Strauss O, Comby F, Aldon MJ (2000) Rough histograms for robust statistics. In: Proceedings of international conference on pattern recognition (ICPR'00), vol II, Barcelona. IEEE Computer Society, pp 2684–2687

Sudkamp T (2000) Similarity as a foundation for possibility. In: Proceedings of 9th IEEE international conference on fuzzy systems, San Antonio, pp 735–740

Yamada K (2001) Probability-possibility transformation based on evidence theory. In: Joint 9th IFSA World Congress and 20th NAFIPS international conference 2001, pp 70–75

Yang Y, Webb GI (2003) Discretization for naive-bayes learning: managing discretization bias and variance. Technical Report 2003-131

Zadeh LA (1978) Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets Syst 1:3–28

Zhang H (2004) The optimality of naive bayes. In: Proceedings of 17th international FLAIRS conference (FLAIRS2004)